INFORMATION AND CONTROL 32, 379–384 (1976)

Estimation of Feature-Dependent Markov Process Transition Probability Matrices*

BRIAN E. BOYLE

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

This article is concerned with the estimation of Markov process transition probabilities for nonhomogeneous populations. This estimation problem is shown to be solvable using general pattern recognition techniques. Numerous multivariate estimation techniques exist in the field of statistical pattern recognition, and many of these will be useful to researchers who use Markov process models of a population's behavior. These techniques are particularly called for when the behavior of members of a population is suspected to depend upon a set of descriptive feature values. One such technique, that of linear discriminant analysis, is presented in detail as an illustration of a statistical pattern recognition approach to a specific Markov process estimation problem. A brief example is given.

1. INTRODUCTION

A number of physical and behavioral processes can be modeled as Markov chains having probability measures associated with state transitions. Parameterization of these Markov chain models requires that transition probabilities from state i to state j, denoted p_{ij}, be estimated for the population as a whole. However, frequently the population is not comprised of a simple homogeneous group, but rather represents the union of several subpopulations. In the more general case, these subpopulations may not be entirely distinct. That is, they may represent a subclassification of the population based upon a partitioning on a set of measurable features.

If $\mathbf{x} = [x_1, x_2, \ldots, x_r]$ represents a vector of feature values describing a member of the population, then in general that member's transition probabilities should be given as a function of its feature vector $\mathbf{x}$. Param-

* This work was supported in part by the National Science Foundation and the MIT Research Laboratory of Electronics. The author is grateful to Professor G. A. Gorry of MIT for several useful discussions of this work.

eterization of this Markov chain requires the estimation of feature-dependent transition probabilities of the form $p_{ij}(\mathbf{x})$. To the best of this author's knowledge, the structure of this estimation problem is not addressed in the published literature of the field. This article will demonstrate that this matrix probability estimation problem is solvable using general pattern recognition techniques. In addition, a linear discriminant technique will be shown to yield a computationally efficient means of estimating the transition probability matrix of a Markov process for which it is known that at most two probabilities in each row are nonzero.

2. Problem Statement

Let $\mathbf{P}(\mathbf{x})$ be the transition probability matrix of an N-state discrete-time Markov process. Assume that $\mathbf{P}(\mathbf{x})$ is first order and stationary for all $\mathbf{x} \in X \subset E^r$. Let $p_{ij}(\mathbf{x})$ be the (i, j)th element of $\mathbf{P}(\mathbf{x})$.

Let $\hat{\mathbf{P}}(\mathbf{x})$ be an estimate of $\mathbf{P}(\mathbf{x})$, with $\hat{p}_{ij}(\mathbf{x})$ an estimate of $p_{ij}(\mathbf{x})$. We wish to develop a technique for obtaining $\mathbf{P}(\mathbf{x})$ from a sample $\{(\mathbf{x}_k, \mathbf{s}_k): k = 1,..., K\}$, where $\mathbf{x}_k$ is the feature vector of the kth sample, and $\mathbf{s}_k$ is the corresponding state occupancy vector; that is, $\mathbf{s}_k$ is a vector whose nth element is the state at which the kth sample was observed to be at time n. The number of elements in $\mathbf{s}_k$ depends on the number of periods the sample was observed. This number may vary from sample to sample.

Let $c_{ij}^{(k)}$ be the number of state i to state j transitions observed for the kth sample. Let

$$c_{ij} = \sum_{k=1}^{K} c_{ij}^{(k)}$$

and

$$c_i = \sum_{j=1}^{N} c_{ij} .$$

Theorem 1. *$p_{ij}(\mathbf{x})$ is independent of $p_{kl}(\mathbf{x})$ for $i \neq k$ and all j and l, $i, j, k, l = 1,..., N$.*

Proof. Since $\mathbf{P}(\mathbf{x})$ is assumed to be first order for all $\mathbf{x}$, its rows are independent.

This result allows us to consider $\mathbf{P}(\mathbf{x})$ row by row, so that instead of a single $N \times N$ estimation problem we have N independent $N \times 1$ estimation problems. We will show below that this problem can be transformed into

N N-class pattern recognition problems. To help make this transformation more clear, we make the notational correspondences

$$\begin{aligned} P_{ij} &\equiv p(\theta_i = j), \\ P_{ij}(\mathbf{x}) &\equiv p(\theta_i = j \mid \mathbf{x}). \end{aligned} \tag{1}$$

That is, the columns of row i are identified with class indices $\theta_i = 1,..., N$. The probability associated with column j of row i then becomes the probability that the row i class index θ_i takes the value j, with this probability a function of the feature vector $\mathbf{x}$. This identification leads to the following result.

THEOREM 2.

$$\begin{aligned} p_{ij}(\mathbf{x}) &\equiv p(\theta_i = j \mid \mathbf{x}) \\ &\propto p(\mathbf{x} \mid \theta_i = j)\, p(\theta_i = j) \\ &\propto p(\mathbf{x} \mid \theta_i = j)\, p_{ij}\,, \end{aligned} \tag{2}$$

where $p(\theta_i = j) \equiv p_{ij}$ is simply the a priori transition probability from state i to state j.

Proof. Theorem 2 follows directly from (1) and Bayes' theorem.

Based on Theorem 2, an estimate for $p_{ij}(\mathbf{x})$ is

$$\hat{p}_{ij}(\mathbf{x}) \propto \hat{p}(\mathbf{x} \mid \theta_i = j)\, \hat{p}_{ij}\,. \tag{3}$$

Given a reasonably large sample we can estimate the a priori transition probabilities p_{ij} by the relative frequency maximum likelihood estimate

$$\hat{p}_{ij} = c_{ij}/c_i\,. \tag{4}$$

The quantity $\hat{p}(\mathbf{x} \mid \theta_i = j)$ is the estimate of a multivariate probability density that can be obtained as follows.

3. OBTAINING $\hat{p}(\mathbf{x} \mid \theta_i = j)$ IN GENERAL

For each sample $(\mathbf{x}_k, \mathbf{s}_k)$, $k = 1,..., K$, let $w_k(i, j)$ be the number of state i to state j transitions given by $\mathbf{s}_k$. Then consider $\mathbf{x}_k$ to be a point in feature space with "weight" $w_k(i, j)$, or equivalently consider there to be $w_k(i, j)$ points in feature space at position $\mathbf{x}_k$. Given this set of weighted samples, $\hat{p}(\mathbf{x} \mid \theta_i = j)$ can be obtained by any of a number of pattern

recognition (multivariate probability density estimation) techniques. One such technique, namely, linear discriminant analysis, will be illustrated below.

4. Estimating $\mathbf{P}(\mathbf{x})$ by Linear Discriminant Analysis

Linear discriminant analysis provides a computationally efficient means of estimating $p(\mathbf{x} \mid \theta_i = j)$ which can be used in the manner described above to estimate $\mathbf{P}(\mathbf{x})$. This technique assumes that for given i and j, $\mathbf{x}$ is normally distributed with mean vector $\boldsymbol{\mu}_{ij}$ and covariance matrix $\boldsymbol{\Sigma}_i$. That is,

$$p(\mathbf{x} \mid \theta_i = j) \propto \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{ij})^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_{ij})]. \tag{5}$$

Given the set of "weighted" samples $\{\mathbf{x}_k, w_k(i,j)\}$ we can obtain estimates $\mathbf{m}_{ij}$ and $\mathbf{S}_i$ for $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\Sigma}_i$, respectively. Given these estimates we can use (5) to compute $\hat{p}(\mathbf{x} \mid \theta_i = j)$ and then use this value in (3) to obtain $\hat{p}_{ij}(\mathbf{x})$.

If the transition matrix $\mathbf{P}(\mathbf{x})$ is known to have at most two nonzero elements in each row, we can further develop this estimation procedure to produce a concise computational formula. Let $j1$ and $j2$ be the nonzero columns of row i. Since the matrix $\mathbf{P}(\mathbf{x})$ is stochastic, $p_{i,j1}(\mathbf{x}) = 1 - p_{i,j2}(\mathbf{x})$.

If we now take the natural logarithm of the ratio of $p_{i,j1}(\mathbf{x})$ to $p_{i,j2}(\mathbf{x})$ we obtain, using (3), (4), and (5) and simplifying,

$$\begin{aligned}\ln[p_{i,j1}(\mathbf{x})/p_{i,j2}(\mathbf{x})] = {} & (\mathbf{m}_{i,j1} - \mathbf{m}_{i,j2})^t \mathbf{S}_i^{-1}\mathbf{x} \\ & - \tfrac{1}{2}(\mathbf{m}_{i,j1} - \mathbf{m}_{i,j2})^t \mathbf{S}_i^{-1}(\mathbf{m}_{i,j1} + \mathbf{m}_{i,j2}) \\ & + \ln[c_{i,j1}/c_{i,j2}].\end{aligned} \tag{6}$$

Defining

$$\mathbf{d}_i{}^t \equiv (\mathbf{m}_{i,j1} - \mathbf{m}_{i,j2})^t \mathbf{S}_i^{-1}, \tag{7}$$

$$b_i \equiv -\tfrac{1}{2}(\mathbf{m}_{i,j1} - \mathbf{m}_{i,j2})^t \mathbf{S}_i^{-1}(\mathbf{m}_{i,j1} + \mathbf{m}_{i,j2}), \tag{8}$$

we obtain

$$\begin{aligned}\ln[p_{i,j1}(\mathbf{x})/p_{i,j2}(\mathbf{x})] &= \mathbf{d}_i{}^t\mathbf{x} + b_i + \ln(c_{i,j1}/c_{i,j2}) \\ &\equiv l_i .\end{aligned} \tag{9}$$

Since $p_{i,j1}(\mathbf{x}) + p_{i,j2}(\mathbf{x}) = 1$, it follows that

$$p_{i,j1}(\mathbf{x}) = \frac{e^{l_i}}{1 + e^{l_i}}; \qquad p_{i,j2}(\mathbf{x}) = \frac{1}{1 + e^{l_i}}. \tag{10}$$

With N states and $\mathbf{x}$ an $(r \times 1)$ vector, a computer program that will estimate these probabilities must store the following numbers of parameters for each row.

for	$\mathbf{d}_i^t$	r parameters
for	$b_i + \ln(c_{i,j1}/c_{i,j2})$	1 parameter
		$r + 1$ total parameters.

Thus storing $N(r + 1)$ parameters and performing slightly more than one matrix multiplication are all that is required to obtain $\hat{\mathbf{P}}(\mathbf{x})$.

To summarize this procedure, $\mathbf{P}(\mathbf{x})$ is estimated row by row. For row i, the set of weighted samples $\{\mathbf{x}_k , w_k(i, j)\}$ is used to obtain the sample mean vectors $\mathbf{m}_{ij}$, $j = j1, j2$, and joint sample covariance matrix $\mathbf{S}_i$. The ith row discriminant vector $(b_i , \mathbf{d}_i^t)$ is then computed from $\mathbf{m}_{i,j1}$, $\mathbf{m}_{i,j2}$, and $\mathbf{S}_i$ using (7) and (8). For any choice of feature vector $\mathbf{x}$ and a priori transition counts $c_{i,j1}$ and $c_{i,j2}$, the transition probability estimates $\hat{p}_{i,j1}(\mathbf{x})$ and $\hat{p}_{i,j2}(\mathbf{x})$ are determined from (9) and (10) by only $r + 1$ multiplications and one exponentiation, where r is the number of features.

5. Example

The payment behavior of an individual using a bank credit card might be modeled as a 5 state Markov process, with states as follows.

state 1: on-time payment,
state 2: 1 month delinquent,
state 3: 2 months delinquent,
state 4: 3 months delinquent,
state 5: defaulted.

A given cardholder might be described by a vector $\mathbf{x} = [x_1 , x_2 , \ldots, x_r]$ of relevant features, for example,

x_1 = annual income,
x_2 = age,
x_3 = years at present residence,

etc.

The state transition probabilities for this cardholder, $p_{ij}(\mathbf{x})$, are expressed as a function of his or her feature vector $\mathbf{x}$. If we had sufficient sample information, we could use the techniques presented in Sections 3 and 4

to estimate the feature-dependent transition probability matrix $\mathbf{P}(\mathbf{x})$. Such a sample might consist of observations on K cardholders with feature vectors $\mathbf{x}_k$ and state occupancy vectors $\mathbf{s}_k$. For example, 9 months of observations on a cardholder whose payments became 1 month delinquent (state 2) in the third month and 3 months delinquent (state 4) in the seventh month could be represented by the vector $\mathbf{s}_k = [1, 1, 2, 1, 2, 3, 4, 1, 1]$. This sample contains two transitions from state 1 to state 2, giving a weight of $w_k(1, 2) = 2$ for the feature vector $\mathbf{x}_k$ in estimating $p_{12}(\mathbf{x})$.

Given such a sample, the feature-dependent transition probability matrix $\mathbf{P}(\mathbf{x})$ can be estimated as indicated above. Such an estimate would be essential to statistical decision making for both initial credit granting and subsequent delinquent collection efforts. A more extensive analysis of this particular example has been conducted (Boyle, 1974).

6. Summary

The estimation of Markov process transition probabilities for nonhomogeneous populations is shown to be solvable using the techniques of statistical pattern recognition. Numerous multivariate estimation techniques exist in the field of statistical pattern recognition, and many of these will be useful to researchers who use Markov process models of a population's behavior. These techniques are particularly called for when the behavior of members of a population is suspected to depend upon a set of descriptive feature values. One such technique, that of linear discriminant analysis, is presented in detail as an illustration of a statistical pattern recognition approach to a specific Markov process estimation problem.

Received: March 14, 1975; revised: February 20, 1976

Reference

Boyle, B. E. (1974), "The Decision to Grant Credit," Ph. D. Dissertation, Massachusetts Institute of Technology.